THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 3, March 1977
Printed in U.S.A.

Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech

By L. R. RABINER, C. E. SCHMIDT, and B. S. ATAL

(Manuscript received September 24, 1976)

Recently, a statistical-decision approach to the problem of voiced-unvoiced-silence detection of speech was proposed by Atal and Rabiner. This method was found to perform well on high-quality speech. However, the five speech parameters used in the analysis were not found to be as good for telephone-quality speech. Thus, an investigation was undertaken to determine a suitable set of parameters that would provide a reliable voiced-unvoiced-silence decision across a variety of standard telephone connections. A large number of parameters (70) were included in the investigation, including 12 LPC coefficients, 12 correlation coefficients, 12 parcor coefficients, 12 LPC partial error terms, etc. Many of the parameters were immediately eliminated because they provided almost no separability between the three decision classes. The remaining parameters were used in a knockout optimization to determine the five best parameters to use for a voiced-unvoiced-silence analysis. Various error weights were investigated to see what types of errors occurred and how they could be minimized. Finally, the use of the Itakura two-pole spectral normalization was investigated to see its effect on the error scores.

I. INTRODUCTION

In a recent paper, Atal and Rabiner described a fairly sophisticated method for reliably classifying segments of a waveform as voiced speech, unvoiced speech, or silence.[1] The analysis method used a statistical pattern-recognition approach to make this three-class decision. In another investigation, Rabiner et al. showed that the accuracy of the classification algorithm was quite high when the input signal was wideband; however, for telephone speech inputs, the accuracy of the classification degraded quite significantly.[2] The reason for this result

was not that the method inherently broke down for telephone inputs, but instead that the particular parameter set effective for wideband inputs was not equally effective for band-limited inputs. Thus, the motivation for the work to be presented in this paper is to investigate the suitability of a large number of parameters as features for reliable voiced-unvoiced-silence classification for telephone-quality speech.

Figure 1 shows a block diagram of the basic voiced-unvoiced-silence analysis algorithm. As shown in this figure, there are three steps in the method. First the speech is preprocessed. Generally, this preprocessing is a simple filtering operation; e.g., in the earlier work, a 200-Hz highpass filter was used to remove dc, hum, or low-frequency noise components present in the input signal. For telephone line inputs, we have considered somewhat more sophisticated preprocessing; namely, we have studied the use of a second-order inverse filter (as originally proposed by Itakura[3]) to normalize out the effects of varying telephone lines.

The second step in the algorithm is the feature measurement stage. For wideband inputs, only five parameters were considered, namely:

(*i*) Energy of the signal
(*ii*) Zero-crossing rate of the signal
(*iii*) Autocorrelation coefficient at unit sample delay
(*iv*) First predictor coefficient
(*v*) Energy of the prediction error.

These measurements were shown to provide a high degree of separability between the three classes of signal for wideband inputs.[1] However, for telephone-quality inputs, the band-limiting of the telephone line considerably reduces the effectiveness of all of the parameters in separating the classes of voiced speech, unvoiced speech, and silence. For example, the absence of signal energy above about 3 kHz significantly reduces the number of zero crossings for unvoiced speech.

To find an effective set of parameters that would be capable of reliably distinguishing between the three signal classes for telephone line inputs, a large number of parameters (70 in total) were studied. Using a set of training data, the probability-density functions for each of the parameters were estimated. Those parameters that provided little or no separation between voiced speech, unvoiced speech, and silence were eliminated from consideration. The remaining 36 parameters were studied as to their effectiveness in classifying telephone line inputs. A knockout type optimization was used to obtain the five most effective parameters for classifying signals according to an error-weighting scheme. Several combinations of different test sets of data and error weights were investigated.

The final step in the analysis method of Fig. 1 is a distance computa-

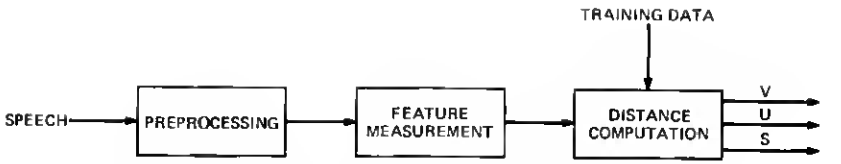

Fig. 1—Block diagram of silence-unvoiced-voiced classification system.

tion to determine whether a test signal is voiced, unvoiced, or silence. For this step, the non-Euclidean distance metric of Ref. 1 was retained because of its invariance properties to linear transformations of the data.[4]

Before presenting the results of the investigation, it is worthwhile reviewing the major distortions of telephone line signals as compared to wideband signals recorded with a high-quality microphone. These distortions include:

(*i*) Band limitation—The frequency response of a telephone line is approximately band limited between 300 Hz and 3000 Hz.

(*ii*) Phase distortion—For the frequency band between 300 and 3000 Hz, the magnitude of the incoming signal remains relatively flat; however, the phase is altered significantly in this band.

(*iii*) Nonlinear effects—Various nonlinearities occur in telephone transmission, including amplitude distortion (signal fading), peak and center clipping, impulse and/or gaussian noise addition, crosstalk, etc.

The effects of the first type of distortion are the most significant as far as this analysis method is concerned.* However, the other types of distortion can, and often do, play a role in determining an effective set of parameters for classifying telephone line signals.

The organization of this paper is as follows. In Section II, we present a description of the techniques used to determine the most effective sets of five parameters for classifying the incoming signals. In Section III, we present the results of the knockout optimization tests for each of the test sets of data and for each set of error weights. Finally, in Section IV, we compare the results on telephone inputs to those obtained with wideband inputs. A typical example showing how the method ultimately performed is presented to illustrate the types of problems that occur with telephone inputs.

* In this work we are considering only those distortions that occur within a local PBX; thus, one would expect a minimum of phase distortion and other nonlinear effects to occur. The place in which such distortions can become significant is in long-distance transmissions.

II. TELEPHONE SIGNAL ANALYSIS SYSTEM

For the preprocessing step of the analysis, a single highpass filter was always used to eliminate hum, dc offset, and low-frequency noise. This filter is described in Ref. 1. A second type of preprocessing was also investigated: the spectrum normalization technique as originally proposed by Itakura.[3] In this technique, the gross long-time spectrum of the signal is estimated using a two-pole LPC model, and then the signal is inverse filtered to remove the gross spectral tilt. Using the two-pole spectral normalization to reduce the spectral variability should, theoretically, also make the feature estimates more reliable. The rationale for considering this form of preprocessing is that for telephone speech the individual telephone transducer and line responses vary greatly across different handsets and telephone lines. Thus, any features estimated over such varying conditions may be adversely affected by the inherent variability of the transmission medium.

The way in which the two-pole spectral normalization was implemented is shown in Fig. 2. For each frame (10 ms of data), three correlation coefficients, $R(m)$, $m = 0,1,2$, are computed using the relation

$$R_j(m) = \sum_{n=0}^{N-m} s_j(n)s_j(n+m) \quad m = 0,1,2, \tag{1}$$

where N is 100, the sampling frequency is 10 kHz, and j is a frame counter that goes from 1 to NF, the number of frames in the utterance. The weighted normalized averages of the first two correlation coefficients (the $m = 1$, $m = 2$ terms) are computed as

$$\overline{R(m)} = \frac{1}{NC} \sum_{j=1}^{NF} \frac{R_j(m)}{R_j(0)} W_j(m), \qquad m = 1,2, \tag{2}$$

where $W_j(m)$ is a weight on the correlation function of the form

$$W_j(m) = \begin{cases} 1 & \text{if } R_j(0) > T \\ 0 & \text{otherwise} \end{cases}; \tag{3}$$

i.e., only frames whose energy $[R_j(0)]$ exceeds a fixed threshold T are

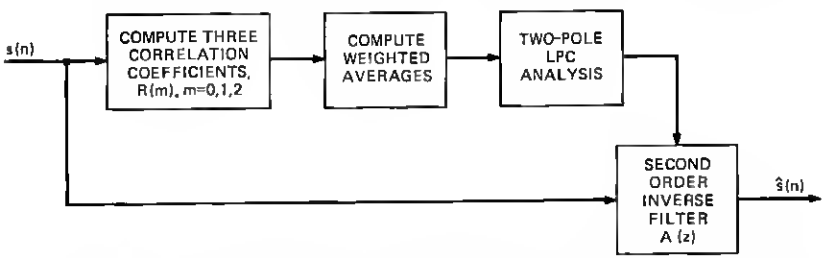

Fig. 2—Block diagram of two-pole spectral-normalization system.

included in the computation of the average correlations. The factor NC in eq (2) is the number of frames that exceeds the threshold of eq. (3). The weighting is used in computing the average correlations to eliminate unvoiced sounds and silence for which the correlation values are significantly different from those for voiced frames.

The third step in the normalization procedure is to compute the predictor coefficients of a two-pole linear predictive coding (LPC) match to the long-time average gross spectrum. If we denote the two LPC coefficients as a_1 and a_2, then the inverse filter needed to normalize the speech spectrum has a transfer function

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2}. \tag{4}$$

On a frame-by-frame basis the inverse filter can be applied directly to the autocorrelation coefficients of a high-order LPC analysis of the signal by convolving them with the autocorrelation coefficients of the second-order inverse filter.[3]

2.1 Features used in the analysis

The parameters (features) studied in the course of this work included the following:

Parameter	Description
1–12	The LPC coefficients of a 12th-order analysis using the Burg lattice method with a 10-ms frame[5,6]: a(1) to a(12).
13–24	The first 12 autocorrelation coefficients of the signal using a 10-ms frame: $\phi(0,1)$ to $\phi(0,12)$.
25–36	The first 12 parcor (partial correlation) coefficients of the signal: $k(1)$ to $k(12)$.
37–48	The first 12 partial normalized error coefficients of the LPC analysis: $E(1)$ to $E(12)$.
49–60	The first 12 cepstral coefficients of the signal as obtained by transforming the LPC coefficients: c(1) to c(12).
61	The log energy of the signal: LE.
62	The number of zero crossings per 10-ms frame: NZ.
63	The log normalized error of the 12-pole LPC analysis: LNE.
64	The maximum value minus the minimum value of the signal during the frame: ML.
65	The absolute energy in the first difference of the signal: ED.
66	The number of zero crossings per 10-ms frame for the first difference signal: NZD.
67	The maximum value minus the minimum value for the first difference signal: MLD.
68	The absolute energy of a smoothed version of the signal: ES.
69	The number of zero crossings per 10-ms frame for the smoothed signal: NZS.
70	The maximum value minus the minimum value for the smoothed signal: MLS.

Figure 3 shows the basic measurement scheme. For each 10-ms frame of the signal, an LPC analysis was performed using the Burg lattice method[5,6] giving a set of 12 LPC coefficients, 12 parcor coefficients, and 12 partial normalized errors defined as

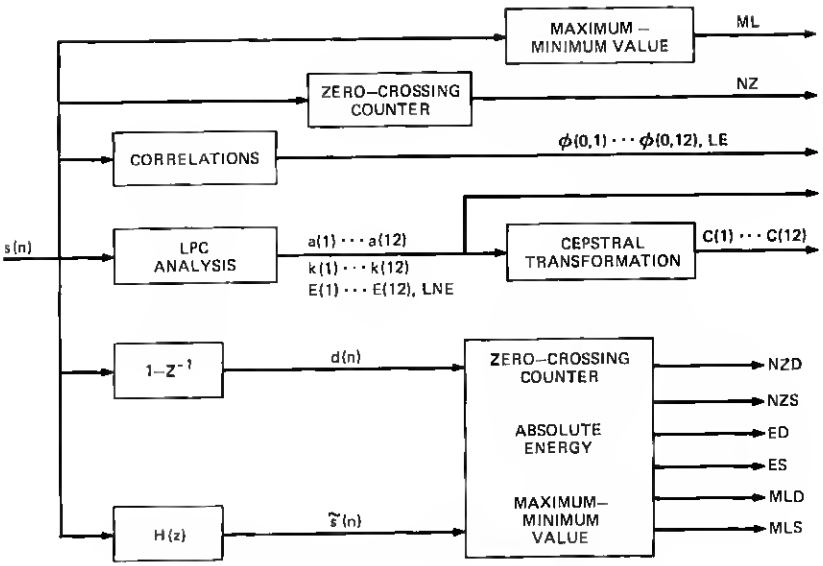

Fig. 3—Block diagram of feature-measurement system.

$$E(l) = \prod_{i=1}^{l} [1-k^2(i)], \qquad l = 1,2,\cdots,12, \tag{5}$$

i.e., $E(l)$ is the normalized error of an l-pole LPC analysis. Since the lattice method does not require the set of correlations directly, they are computed on the signal from the equation

$$\phi(o,i) = \sum_{n=0}^{N-1} s(n)s(n-i), \qquad i = 1,2,\cdots,12, \tag{6}$$

i.e., a nonstationary correlation function is computed. The cepstral coefficients are computed directly from the LPC coefficients using the recursion relation

$$c(i) = a(i) - \sum_{m=1}^{i-1} \frac{m}{i} c(m)a(i-m), \qquad 1 \leqq i \leqq 12. \tag{7}$$

Two other measurements are made directly on the signal $s(n)$. These are the zero-crossing count defined as the number of zero crossings per 10-ms interval, and a computation of the difference between the maximum and minimum signal amplitudes in tbe frame.

In addition to the above parameters, six additional measurements are made on the first difference of the signal, $d(n)$, defined as

$$d(n) = s(n) - s(n-1) \tag{8}$$

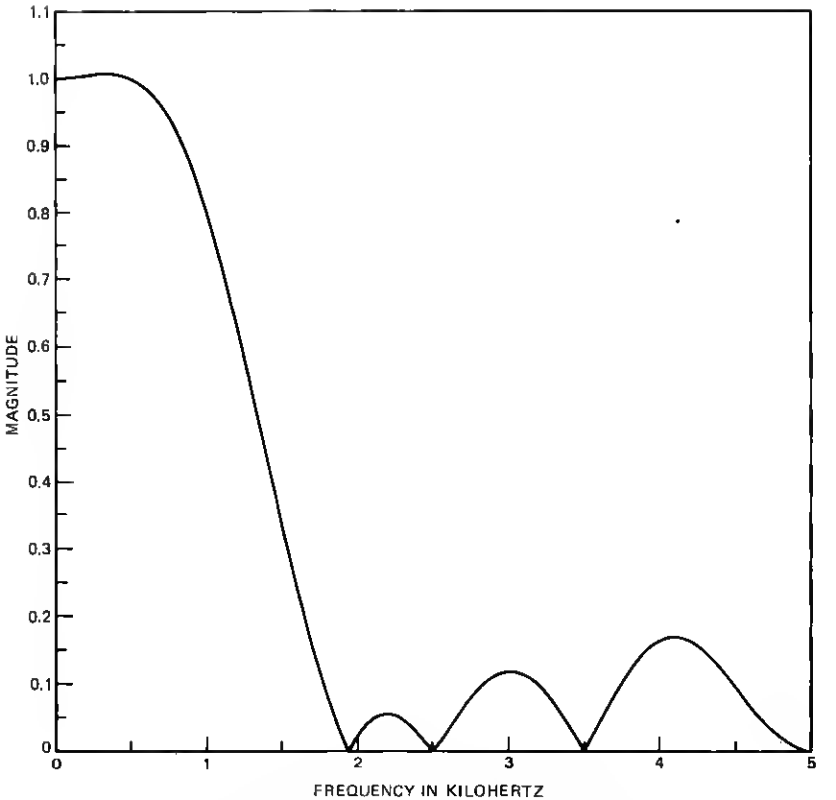

Fig. 4—Frequency response of lowpass smoothing filter.

and a smoothed version of the signal obtained via the filtering relation*

$$\tilde{s}(n) = -s(n) + s(n-2) + 2s(n-3) + 4s(n-4) + 4s(n-5) + 4s(n-6) + 2s(n-7) + s(n-8) - s(n-10). \quad (9)$$

It can be seen that the filtering of eq (9) can be accomplished without the need for a multiplier and, thus, can be implemented quite efficiently. Figure 4 shows the frequency response of this filter. It can be seen that the filter provides a small amount of high-frequency attenuation and therefore can be considered as a lowpass smoothing filter. The measurements made on $d(n)$ and $\tilde{s}(n)$ are zero-crossing count, absolute energy, and difference between maximum and minimum signal levels in the frame.

* This filter as well as parameters 65–70 were suggested by D. R. Reddy for inclusion in this work.

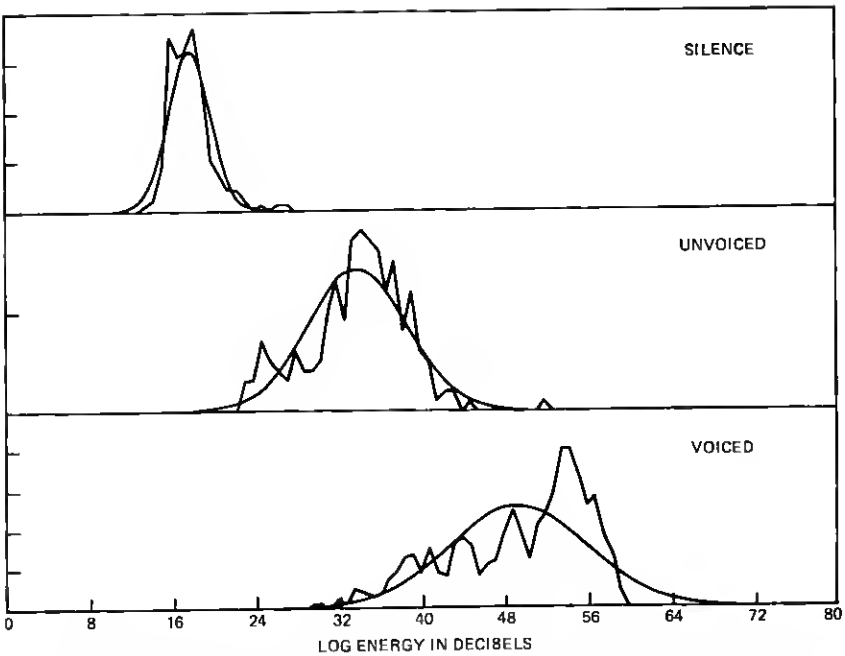

Fig. 5—Probability distributions for log energy of the signal for silence, unvoiced, and voiced classes. Both estimated and gaussian fits to the distributions are shown.

Once the initial set of 70 parameters was chosen, a training set of data was used to estimate one-dimensional probability functions for each of the parameters and for each signal classification. A one-dimensional gaussian curve having the same mean and standard deviation as the measured distributions was also computed for each parameter.* Figures 5 through 7 show three typical distributions for the parameters log energy (feature 61), first LPC coefficient (feature 1), and twelfth LPC coefficient (feature 12), respectively. For the log-energy parameter (Fig. 5), the distributions for silence, unvoiced, and voiced speech were fairly well separated with means of 18, 34, and 49 dB, respectively. Similarly the distributions for the first LPC coefficient (Fig. 6) were also well separated with means of −0.19, −0.66, and −1.9 for silence, unvoiced, and voiced speech, respectively. However, as shown in Fig. 7, the distributions for all parameters were not well separated across the different classes. In this case, the distributions for all three signal classes overlapped considerably. It seems reasonable that features in which such behavior is observed will not be effective in the classification procedure. Therefore,

* For the distance metric used in this work, it is not critical that the one-dimensional distributions of the parameters be well approximated by a simple gaussian curve. It is important, however, that the distributions be unimodal.

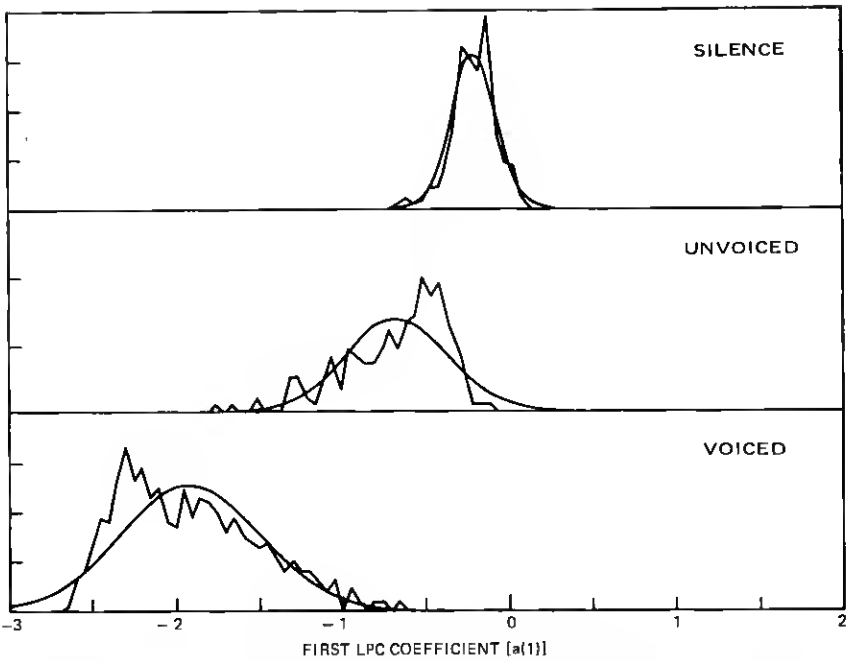

Fig. 6—Probability distributions for first LPC coefficient for silence, unvoiced, and voiced classes.

such parameters were not considered in the testing to be described in this paper.

A total of 34 of the 70 parameters were eliminated in this manner. The parameters eliminated were the higher LPC coefficients [a(5) to a(12)], the higher autocorrelation coefficients [$\phi(0,5)$ to $\phi(0,12)$], the higher parcor coefficients [$k(5)$ to $k(12)$], the higher cepstral coefficients [$c(5)$ to $c(12)$], and the last two partial normalized LPC error coefficients [$E(11)$ and $E(12)$]. The remaining 36 parameters were used in all the optimization tests described in the next section.

2.2 Knockout optimization procedure

To choose the set of five parameters out of the remaining 36 features that best (most accurately) classified signal intervals as silence, unvoiced, or voiced speech, a knockout optimization procedure was used.[7] Figure 8 shows a flow diagram of the procedure. Using a testing set of data (see Section 2.3) and an objective error measure, the knockout optimization proceeded first to find the single best parameter for separating the three classes. The best parameter is knocked out and used in combination with each of the remaining 35 features to find the best pair of parameters for the signal classification. This process of knocking out the best parameter

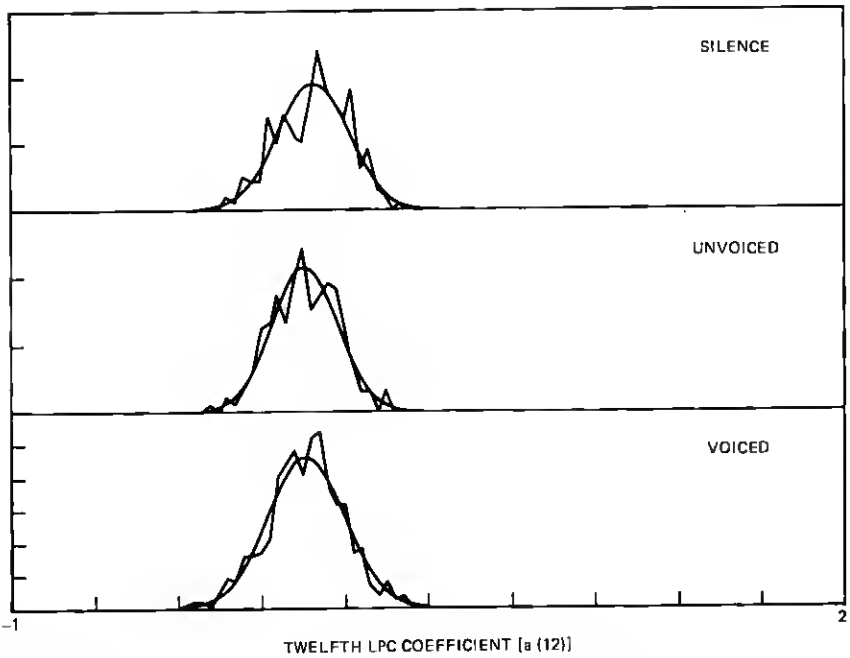

Fig. 7—Probability distributions for twelfth LPC coefficient for silence, unvoiced, and voiced classes.

and combining all knocked-out features with the ones remaining in the parameter set was iterated until a total of five parameters were obtained.

Several comments should be made about this procedure. First, it is noted that this method does *not* necessarily yield the optimum set of five parameters for making the silence, unvoiced, voiced decision. In general, the resulting parameter set is suboptimal since only a very small subset of the total number of combinations of 36 parameters taken five at a time are considered in this method. In defense of the method, however, one can argue that, within the constraints of the procedure, an optimal set of the 36 parameters is chosen. Furthermore, at least theoretically, the addition of each new knocked-out feature reduces the error score. Finally, it is argued that the resulting feature sets provide significantly better accuracy for signal classification than almost any randomly chosen set of five of the 36 parameters.

2.3 Distance computation

The distance computation used throughout this investigation was the non-Euclidean distance metric defined in Ref. 1. For the feature vector $\mathbf{x} = [\mathbf{x}(1),\mathbf{x}(2),\ldots,\mathbf{x}(5)]$ with mean vector $\mathbf{m}_i = [\mathbf{m}_i(1),\mathbf{m}_i(2),\ldots,\mathbf{m}_i(5)]$

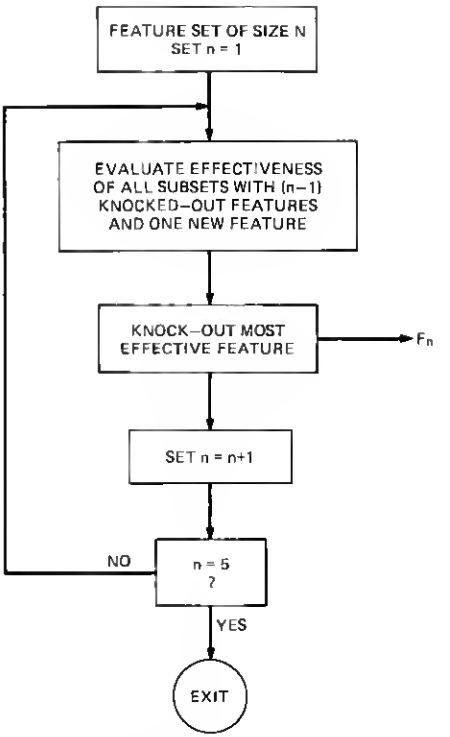

Fig. 8—Flow chart of knockout optimization algorithm.

and covariance matrix Λ_i, the distance computation was of the form

$$d_i = (\mathbf{x} - \mathbf{m}_i)\Lambda_i^{-1}(\mathbf{x} - \mathbf{m}_i)^t, \tag{10}$$

where $i = 1$ (silence), 2 (unvoiced), or 3 (voiced), and t denotes the transpose of a vector. For each signal class, d_i is computed and the decision rule is to select class i such that $d_i < d_j$ for all $j \neq i$; i.e., choose the class with the minimum distance to vector **x**.

To implement the distance computation in eq. (10) during the knockout optimization required the computation of a new covariance matrix Λ_i for each subset of parameters being considered. Thus, on the order of 420 covariance matrices had to be estimated in a typical optimization run. This represented a substantial amount of computation.

2.4 Experimental procedure

The formal evaluation of the feature sets was made by choosing a fairly large data base of different utterances and different speakers, using part of the data base for training the system, and using the remainder of the data base for testing the system.

A total of five speakers (two male, three female) were used in the telephone-line evaluation. Each speaker recited three utterances,* each one over a new dialed-up connection and thereby ensuring that a different telephone-transmission path was obtained for each utterance. Two different telephone transmitters (carbon microphones) were also used in the test. One utterance from each speaker was used in the training set; the remaining two utterances were used in the testing set.

For each recorded utterance, a manual analysis was performed on each 10-ms interval to classify it as voiced, unvoiced, or silence based on both the acoustic waveform and a phonetic transcription of the utterance. Each signal classification was further modified with a label as to the certainty of the manual classification. The labels used were:

(*i*) Absolutely certain—clear characteristics of the class to which it was assigned.

(*ii*) Moderately certain—generally a boundary interval between classes in which two types of signal were present.

(*iii*) Uncertain—classified primarily on linguistic information about the utterance. Included in this class were voiced fricatives, voiced stops, and certain transients (including some telephone-line transients).

Figure 9 shows an example in which uncertain intervals occurred. This section of speech is from the beginning of the word *cowboys*. The initial intervals should linguistically be labelled as either silence or unvoiced speech corresponding to the stopgap and burst of the voiceless stop /k/. However, acoustically the initial seven intervals (as marked in Fig. 9) show properties more similar to voiced speech than to silence or unvoiced sounds. These intervals were treated as uncertain intervals and were marked as unvoiced speech for testing purposes.

For the training set, only those intervals for which the classification was absolutely certain were used. For the testing set, three sets of data were used. One set contained only those intervals for which the classification was absolutely certain (TS1). The second set contained both the moderately certain as well as the absolutely certain intervals (TS2). The third set contained all the intervals, regardless of the certainty of manual classification (TS3). In the next section, we present results for each of these testing sets of data.

III. RESULTS

The knockout optimization procedure described in Section II was run on the three sets of test data using 10 different error-weighting matrices.

* Each utterance was a carefully chosen sentence containing a mixture of voiced, unvoiced, and silence intervals.

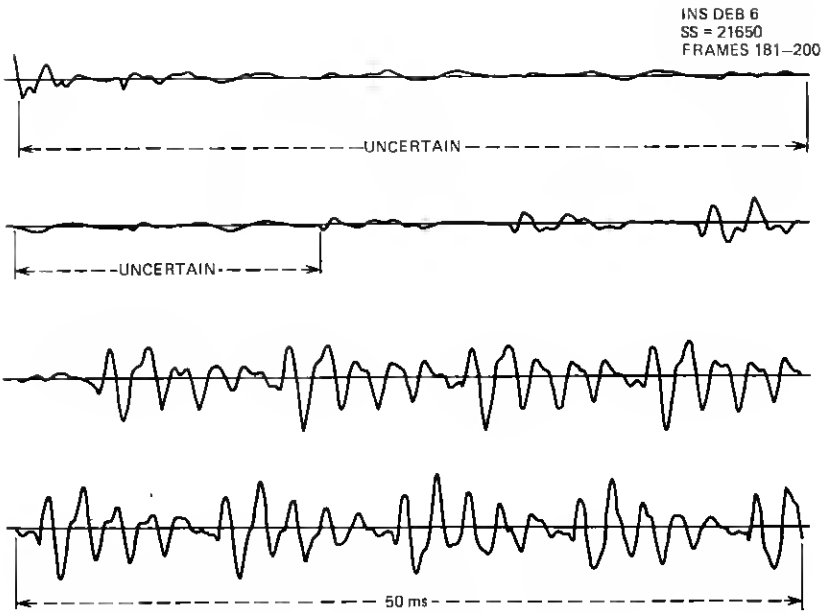

Fig. 9—The acoustic waveform for a section of speech in which an interval was uncertain.

In addition, the entire experiment was rerun on data preprocessed using the two-pole spectral normalization method described in Section II. Table I provides a summary of the three test sets of data, the 10 error-weight matrices, and the two processing conditions.

The error-weight matrices were used to study the effects of weights for each type of classification error on the overall error rate and the choice of the optimal features. The definition of a general error-weight matrix is as follows. If we let E denote the overall error score in classifying the data of a test set, then

$$E = N_{ss}W_{ss} + N_{su}W_{su} + N_{sv}W_{sv} + N_{us}W_{us} + N_{uu}W_{uu} + N_{uv}W_{uv} + N_{vs}W_{vs} + N_{vu}W_{vu} + N_{vv}W_{vv}, \quad (11)$$

where N_{ab} is the number of frames of a class a which were classified as belonging to class b, and W_{ab} is the weight attached to this pair of classifications. It should be clear from eq. (11) that

$$\begin{aligned} N_s &= N_{ss} + N_{su} + N_{sv} \\ N_u &= N_{us} + N_{uu} + N_{uv} \\ N_v &= N_{vs} + N_{vu} + N_{vv}, \end{aligned} \quad (12)$$

where N_a is the number of frames in the test set in class a. Table II shows

Table I — Summary of factors considered in the investigation

Data Test Sets	TS1	Absolutely certain intervals
	TS2	Moderately certain intervals added to TS1
	TS3	Uncertain intervals added to TS2
Error-Weight Matrices	WM1	Uniform matrix
	WM2	Silence weighting matrix
	WM3	Unvoiced weighting matrix
	WM4	Voiced weighting matrix
	WM5	Silence-to-unvoiced weighting matrix
	WM6	Unvoiced-to-silence weighting matrix
	WM7	Silence-to-voiced weighting matrix
	WM8	Voiced-to-silence weighting matrix
	WM9	Voiced-to-unvoiced weighting matrix
	WM10	Unvoiced-to-voiced weighting matrix
Preprocessing	P1	Direct transmission
	P2	Two-pole spectral normalization

the 10 weight matrices described in Table I. Each matrix is expressed in the form

$$W = \begin{bmatrix} W_{ss} & W_{su} & W_{sv} \\ W_{us} & W_{uu} & W_{uv} \\ W_{vs} & W_{vu} & W_{vv} \end{bmatrix}, \tag{13}$$

where W_{ab} is not generally the same as W_{ba}.

As seen in Table II, error weight-matrix 1 (WM1) attaches equal weight to all six types of misclassifications and, therefore, is the canonic error matrix for the three-class problem. Error matrices 2–4 (WM2–WM4) each choose a subset in which one of the three classes is essentially merged with another class. For example, error matrix 4 (WM4) gives 0 weight to errors between the classes of silence and unvoiced speech; however, the other four types of error have unity weight. Thus, this matrix serves to distinguish most effectively between voiced speech and nonvoiced (either silence or unvoiced) speech. As another example, error matrix 2 (WM2) gives 0 weight to errors between the classes of voiced and unvoiced speech. Thus, this matrix serves to distinguish between speech (voiced or unvoiced) and silence. As such, it would be useful for speech-detection applications. Error matrices 5 through 10 each focus on only one of the six sets of misclassifications. The results for these cases give a lower bound on the error rate for special cases in which only a single type of misclassification is considered.

For each of the sets of data of Table I, the knockout optimization procedure was used giving the five best features and the resulting overall misclassification rate, defined as

$$E_N = \frac{E}{(N_s + N_u + N_v)}. \tag{14}$$

Table II — Error-weight matrices used in the investigation

$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ WM1 (a)	$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ WM2 (b)	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ WM3 (c)
$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ WM4 (d)	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ WM5 (e)	$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ WM6 (f)
$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ WM7 (g)	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ WM8 (h)	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ WM9 (i)
$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ WM10 (j)		

For error weights 5 through 10 (where only a single misclassification was counted) the overall misclassification rated was defined as

$$E_N = \frac{E}{N_a}, \tag{15}$$

where

$$N_a = \begin{cases} N_s & \text{for WM5 and WM7} \\ N_u & \text{for WM6 and WM10} \\ N_v & \text{for WM8 and WM9} \end{cases}. \tag{16}$$

The results of these experiments are presented in Tables III through VI. Tables IV through VI present the misclassification rate results, and Table III gives both the parameter numbers and the mnemonics of the five parameters chosen by the optimization procedure. The results in these tables are presented sequentially; i.e., the results obtained using only l of the five parameters (l = 1,2,3,4) are indicated in the appropriate rows of the tables.

Two comments should be made about the data. In many cases, it was found that the overall misclassification rate did not monotonically decrease as more features were knocked out of the parameter set. For these

Table III — Optimal features chosen by the knockout optimization for telephone inputs

Test Set	Parameter	Weight Matrix									
		WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS1	1	61-LE	65-ED	14-ϕ(0,2)	63-LNE	65-ED	70-MLS*	67-MLD	68-ES*	63-LNE	46-E(10)
	2	14-ϕ(0,2)	26-k(2)	68-ES	44-E(8)	27-k(3)				2-a(2)	42-E(6)
	3	63-LNE	16-ϕ(0,4)	69-NZS	68-ES	67-MLD*				68-ES	43-E(7)*
	4	67-MLD			4-a(4)					65-ED	
	5	64-ML								27-k(3)*	
TS2	1	61-LE	65-ED	14-ϕ(0,2)	63-LNE	65-ED	64-ML	67-MLD	68-ES*	63-LNE	46-E(10)
	2	50-c(2)	26-k(2)	68-ES	66-NZD	27-k(3)	70-MLS*			2-a(2)	42-E(6)
	3	16-ϕ(0,4)		69-NZS	45-E(9)	67-MLD*				68-ES	62-NZ
	4				70-MLS					65-ED	25-k(1)
	5				67-MLD					27-k(3)*	45-E(9)*
TS3	1	61-LE	65-ED	14-ϕ(0,2)	63-LNE	65-ED	50-c(2)	67-MLD	68-ES*	14-ϕ(0,2)	46-E(10)
	2	50-c(2)	26-k(2)	68-ES	66-NZD	27-k(3)	52-c(4)		43-E(7)*	70-MLS	42-E(6)
	3	16-ϕ(0,4)		69-NZS	45-E(9)	67-MLD*	68-ES*			63-LNE	40-E(4)
	4	3-a(3)			70-MLS					68-ES	43-E(7)
	5	64-ML			61-LE						66-NZD

* Other features provided the same overall misclassification rate.

Table IV — Error rates for telephone inputs

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
				TS1 Without Two-Pole Spectral Normalization						
1	15.1	4.7	9.8	2.8	2.4	0.7	2.4	0	0.9	1.0
2	11.0	4.7	5.7	2.3	0.6	0			0.5	0.7
3	7.5	4.4	5.3	1.9	0				0.4	0.3
4	6.9			1.9					0.3	
5	6.4									

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
				TS1 With Two-Pole Spectral Normalization						
1	13.8	5.1	11.0	5.1	2.4	4.6	2.4	0	3.2	6.8
2	9.7		7.1	4.5		0.7			3.2	2.7
3	8.6		6.0	4.4					2.1	2.3
4	7.6		6.0	3.9					1.9	
5	7.6								1.8	

Size of Training and Testing Sets for TS1

	Number of Frames	
	Training	Testing
S	207	328
U	210	306
V	539	1180
	956	1834

Table V—Error Rates for Telephone Line Inputs

Parameter	Weight Matrix									
	WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS2 without two-pole Spectral Normalization										
1	16.2	5.8	12.2	4.5	2.4	0.5	2.6	0	1.8	2.7
2	11.7	5.4	7.4	4.3	0.8	0			1.0	2.1
3	10.8		7.1	3.6	0				0.7	1.3
4				3.5					0.5	1.1
5				3.5						

Parameter	Weight Matrix									
	WM1	WM2	NM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS2 with two-pole Spectral Normalization										
1	18.1	7.2	14.4	8.6	2.9	3.9	2.7	0.3	4.5	9.5
2	13.4		10.4	8.2		1.1		0	5.2	6.1
3	12.4		9.4	7.8					3.9	5.0
4	11.6			7.4					3.8	4.2
5	11.5			7.2					3.6	3.7

Size of Training and Testing Sets for TS2

	Number of Frames	
	Training	Testing
S	207	375
U	210	378
V	539	1196
	956	1949

cases data are presented up to the number of parameters at which the error rate kept decreasing. The second comment concerns the specific features knocked out in the optimization (as given in Table III). In many cases, a large number of features (other than the ones presented) provided essentially the same overall misclassification rate as the feature that was knocked out. These cases are indicated by an asterisk after the feature number in these tables. For such cases, features other than the ones indicated in the table may be equally appropriate.

IV. ANALYSIS OF THE RESULTS

Several important observations can be made by examining carefully the results of Tables III though VI. First, it can be seen by comparing error rates for matrix WM1 to those for matrices WM2 through WM4 that most of the overall error rate for the canonic error matrix was due to misclassifications between the classes of silence and unvoiced speech* (compare results for WM1 and WM3). This result is certainly not unan-

* Further evidence of this result is given in Table VII, which shows a breakdown of the error components. This table is discussed later in this section.

Table VI — Error rates for telephone line inputs

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS3 Without Two-Pole Spectral Normalization										
1	18.5	6.3	13.4	6.3	2.4	0.9	2.7	0	3.3	5.9
2	13.2	5.6	9.0	6.1	0.8	0.2			2.1	5.4
3	12.2		8.7	6.1	0	0			1.3	5.0
4	12.0		8.7	5.2					1.1	4.1
5	11.7			5.0						3.4

Parameter	Weight Matrix WM1	WM2	WM3	WM4	WM5	WM6	WM7	WM8	WM9	WM10
TS3 With Two-Pole Spectral Normalization										
1	20.3	8.2	16.2	10.3	2.9	4.5	2.9	0.2	4.9	12.5
2	15.6		12.2	9.5		1.3		0	5.9	14.3
3	14.3		11.2	9.1					4.8	11.4
4	13.7			8.7					4.5	9.8
5	13.4			8.5					4.4	8.7

Size of Training and Testing Sets for TS3

	Number of Frames: Training	Testing
S	207	379
U	210	443
V	539	1264
	956	2086

ticipated since the band limiting of telephone speech has the most severe effect on unvoiced sounds whose spectral components often fall above the high-frequency cutoff of the telephone transmission.

Based on the above result it would seem reasonable to compare error scores using error matrices 1 and 4. It can be seen from the tables that if one does not consider distinctions between silence and unvoiced speech, then an improvement of somewhat more than 2 to 1 in error score is obtained. For the case of absolutely certain classifications, an error rate of 1.9 percent is obtained for error matrix 4. For test sets TS2 and TS3, the error rate for error matrix 4 increases to 3.5 and 5.0 percent, respectively.

The results using error matrix 2 (the speech detection matrix) show that, in the case of absolutely certain classifications (Table IV) an error rate of 4.3 percent is obtained. For test sets TS2 and TS3, the error rate for matrix 2 increases slightly to 5.4 and 5.6 percent, respectively.

The results of using error matrices WM5–WM10 show that the most frequent misclassification occurs between silence and voiced speech in which error rates on the order of 2 to 3 percent were obtained for all three test cases. The problem here occurs during low-level sounds, such as voiced stops where the silence regions are often classified as voiced due to the presence of low-frequency components of the signal. Unfortunately, such signals do not fall neatly into either category and the decision algorithm consistently classified them as voiced sounds whereas the manual classification was silence.

Comparisons of the results of Tables IV, V, and VI showed that the error scores increased with the complexity of the test set as anticipated. However, it is difficult to attach too much meaning to the absolute error rates for TS2 and TS3, since the frames which were added constituted boundary frames and frames which were subject to classification error in the manual classification. The results are presented to provide information as to the sizes of the increases in error rate that are to be expected with such input test sets.

The data of Table III (the optimal feature list) are also quite interesting. The influence of the weight matrix is evident by scanning across the rows of the table. Each weight matrix had its own set of optimal features, which were different from those of any other weight matrix. By scanning down the columns of this table, however, it is seen that the influence of the data test set was fairly weak in that the optimal-feature set remained substantially the same for all three test sets across the three sets of data.

An interesting result shown in Tables IV through VI is that the two-pole normalization scheme did not provide essentially any improvement in the classification accuracy across any of the test conditions studied. This result is a little surprising in light of the work of Itakura who found

that it compensated different telephone transmission conditions quite adequately.[3] One possible reason for this result is that the non-Euclidean distance metric to some extent compensates automatically for the variable telephone transmission conditions by appropriate linear transformation of the feature space. Thus, for this classification method, the use of a two-pole spectral normalization is of little value.

An additional breakdown of the error analysis for the most important error-weight matrices (WM1, WM2, and WM4) is given in Table VII in which the percentage of each type of misclassification is presented. It can be seen in this table that certain types of errors dominated the scores. For example, no cases occurred throughout the test in which a voiced interval was classified as silence. It can also be seen that, as mentioned previously, the error rate for silence-to-unvoiced speech dominated the overall error rates for error matrices WM1 and WM2, whereas no single component of the error dominated the overall error rate for matrix WM4.

4.1 Comparison with wideband results

Although some numerical scores were presented in Ref. 1 for misclassification rates using the analysis method on wideband (high-quality) data, a set of companion results were obtained in this study to compare and contrast the error results for wideband and telephone signals. Using the identical procedures discussed in Sections II and III, a set of optimal features and error rates were obtained for wideband test sets of signals. The results of these runs are presented in Tables VIII and IX. Comparisons of the error rate tables (VII and VIII) show the following:

(*i*) For error weight matrix WM1, the scores for wideband data were from two to three times lower than for telephone data. This is due to the vastly improved scores on the category of silence-to-unvoiced errors. The error rates for many of the other possible misclassifications were quite comparable.

(*ii*) For error weight WM4, the scores for wideband data were only slightly better than for telephone data, indicating that a voiced-not voiced decision can be as reliably made over a telephone line as for high-quality inputs. However, the speech–not speech decision is much more difficult for telephone data than for wideband signals.

(*iii*) For error weight matrix WM2, the scores for wideband data were from two to eight times lower than for telephone data. This result is again due to the improved performance in discriminating between silence and unvoiced speech for wideband data.

Table VII — Breakdown of error percentages for telephone line inputs

Test Set	SU	SV	US	UV	VS	VU	S	U	V	Overall
	Error-Weight Matrix WM1									
TS1	16.8	6.1	4.6	5.2	0	0.9	22.9	9.9	0.9	6.4
TS2	30.1	5.9	2.9	5.0	0	4.1	36.0	8.0	4.1	11.0
TS3	28.8	5.8	2.9	15.4	0	2.5	34.6	18.3	2.5	11.7
	Error-Weight Matrix WM4									
TS1	93.0	4.6	0	3.0	0	0.9	97.6	3.0	0.9	1.9
TS2	88.9	6.7	0.3	6.1	0	1.6	95.7	6.4	1.6	3.5
TS3	43.1	6.1	3.2	8.6	0	3.4	49.2	11.8	3.4	5.0
	Error-Weight Matrix WM2									
TS1	11.6	5.2	7.6	17.5	0.1	4.7	16.8	25.1	4.8	4.4
TS2	14.8	5.4	7.5	27.7	0.2	7.0	20.2	35.2	7.2	5.4
TS3	15.4	5.6	7.9	27.6	0.2	9.4	21.0	35.5	9.5	5.6

* These results had 0 weight in the overall error score and, therefore, did not affect the choice of features.
** Only a single component of this error score is included in the overall score.

Table VIII — Breakdown of error percentages for wideband inputs

	Error-Weight Matrix WM1									
Test Set	SU	SV	US	UV	VS	VU	S	U	V	Overall
TS1	2.3	0	3.9	2.6	0	1.5	2.3	6.6	1.5	2.2
TS2	8.7	5.8	5.7	6.7	0.1	2.3	14.5	12.4	2.4	5.3
TS3	8.9	5.9	7.2	6.8	0.1	2.3	14.8	14.0	2.5	5.7
	Error-Weight Matrix WM4									
Test Set	SU*	SV	US*	UV	VS	VU	S**	U**	V	Overall
TS1	29.5	0	5.3	2.5	0	0.9	29.5	7.9	0.9	1.1
TS2	39.9	5.8	4.8	9.1	0	1.4	45.7	13.9	1.4	3.1
TS3	33.3	5.9	5.0	5.9	0	2.0	39.3	10.9	2.0	3.1
	Error-Weight Matrix WM2									
Test Set	SU	SV	US	UV*	VS	VU*	S	U**	V**	Overall
TS1	3.4	0	2.0	2.0	0	2.0	3.4	4.0	2.0	0.5
TS2	7.3	7.3	4.3	11.0	0	2.7	14.5	15.3	2.7	2.3
TS3	7.4	7.4	5.9	11.3	0	3.0	14.8	17.2	3.0	2.6

* These results had 0 weight in the overall error score and therefore did not affect the choice of features.
** Only a single component of this error score is included in the overall score.

Table IX — Optimal features for wideband test sets and error-weight matrices WM1, WM4, and WM2

Test Set	WM1	WM4	WM2
TS1	13 $\phi(0,1)$	68 ES	64 ML
	70 MLS	25 k(1)	70 MLS
	14 $\phi(0,2)$	69* NZS	61 LE
	61 LE	67* MLD	14 $\phi(0,2)$
	68* ES		68 ES
TS2	13 $\phi(0,1)$	68 ES	64 ML
	70 MLS	25 k(1)	70 MLS
	14 $\phi(0,2)$	16 $\phi(0,4)$	61 LE
	61 LE	64 ML	68 ES
	68 ES		67 MLD
TS3	13 $\phi(0,1)$	68 ES	64 ML
	70 MLS	15 $\phi(0,3)$	70 MLS
	14 $\phi(0,2)$	26 k(2)	61 LE
	61 LE	13 $\phi(0,1)$	68 ES
	68 ES	52 c(4)	67 MLD

4.2 Typical test example

Figures 10 and 11 show the results of applying the classification method to the utterance, "Few thieves are never sent to the jug," spoken by a male speaker. The contour shown in (a) of each figure is a manual classification of each frame. Part (b) shows the results of analysis using parameters obtained from WM1 (Fig. 10) and WM4 (Fig. 11). Part (c) shows the results of nonlinearly smoothing the analysis contours using a median smoother.[8] Parts (d), (e), and (f) show plots of the probability of correct classification based on the distance calculation for each class; i.e., if we denote the distance calculated for silence as D_s, the distance calculated for unvoiced as D_u, and the distance calculated for voiced as D_v, then

$$P(S) = \frac{D_u D_v}{D_s D_u + D_s D_v + D_u D_v} \tag{17}$$

$$P(U) = \frac{D_s D_v}{D_s D_u + D_s D_v + D_u D_v} \tag{18}$$

$$P(V) = \frac{D_s D_u}{D_s D_u + D_s D_v + D_u D_v}. \tag{19}$$

It can be seen that $P(S)$, $P(U)$, and $P(V)$ define a probability measure, since

$$P(s) + P(u) + P(v) = 1 \tag{20}$$

and

$$0 \leqq P(s), P(u), P(v) \leqq 1 \tag{21}$$

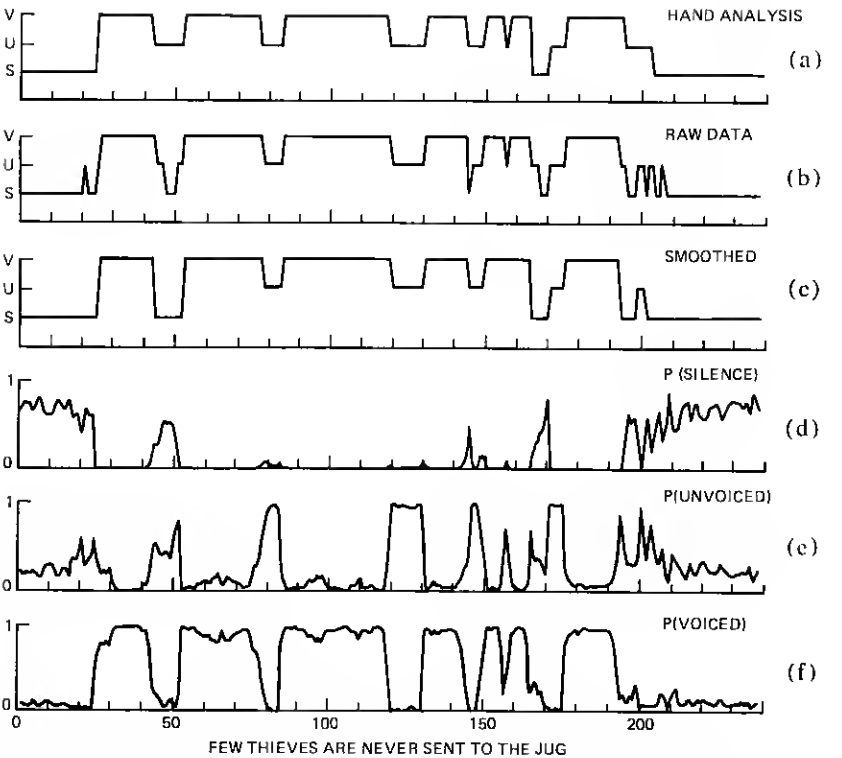

Fig. 10—The analysis results for the utterance, "Few Thieves are Never Sent to the Jug," using optimal features from TS1 with weight matrix WM1.

for all values of D_s, D_u, and D_v. Furthermore, the probabilities satisfy the relation

$$\lim_{Da \to 0} P(a) \to 1 \tag{22}$$

and

$$\lim_{Da \to \infty} P(a) \to 0. \tag{23}$$

Thus, as the distance increases, the probability measures decrease.

Contrasting the silence-unvoiced-voiced contours of Figs. 10 and 11, the following observations can be made:

(*i*) The results obtained using features derived from matrix WM4 essentially never classified frames as silence. Instead all silence frames

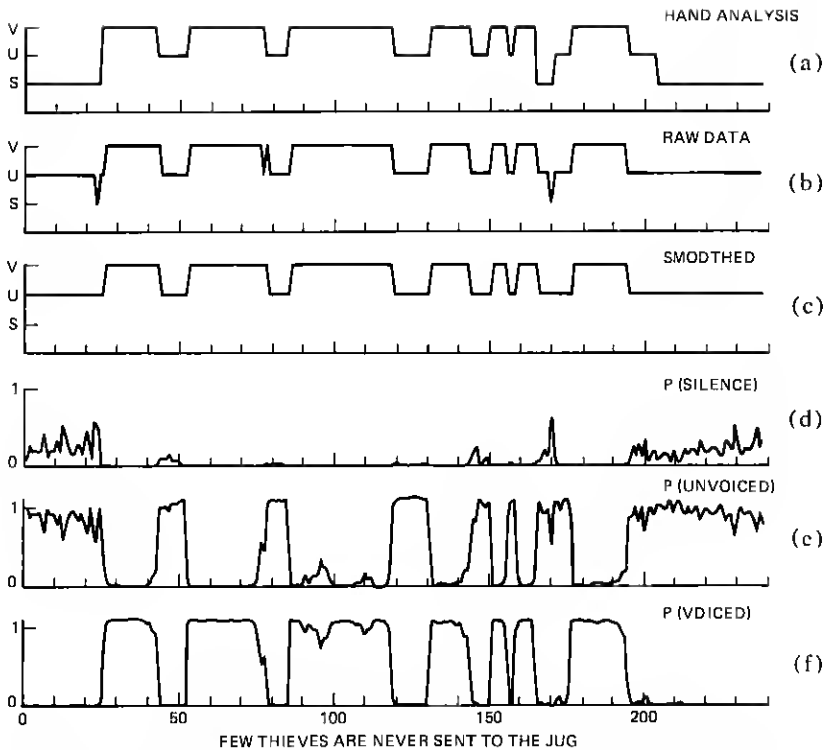

Fig. 11—The analysis results for the same utterance as Fig. 10 using optimal features from TS1 with weight matrix WM4.

were classified as unvoiced, consistent with the zero weight given to this type of error.

(*ii*) Both sets of results contain only a small number of misclassifications of voiced intervals. All but one of these voiced misclassifications occurred at boundaries between voiced and nonvoiced speech.

(*iii*) The probability measures for voiced speech using features derived from WM4 were somewhat higher throughout the voiced regions than corresponding results derived from WM1 features. This indicates that a somewhat better feature set for voiced sounds is obtained at the tradeoff of the high error rate for silence-to-unvoiced errors (and vice versa).

Results similar to those discussed above have been obtained for a wide variety of utterances tested on the system using these sets of features. It is concluded that if one is willing to forego any attempt at distin-

guishing between unvoiced sounds and silence, then reliable voiced-nonvoiced decisions can be obtained over telephone lines.

V. SUMMARY

Through a series of fairly extensive tests, we have investigated quite thoroughly the potential of a fairly sophisticated silence-unvoiced-voiced classification system. We have shown that, depending on the weight attached to various types of misclassifications, a set of optimal features can be found that minimizes the weighted misclassification error rate. For telephone line inputs, the results showed that reliable discrimination between silence and unvoiced sounds is quite difficult; however, reliable discrimination between voiced and nonvoiced sounds (silence or unvoiced speech) can be achieved at error rates fairly close to those obtained with wideband input signals.

Extensive testing of the optimal feature sets obtained from the analysis showed the method to be reliable enough for use in several typical applications in the area of man-machine communication by voice.[9,10]

One aspect of the analysis system which was not varied was the distance metric used in the final classification. Although the non-Euclidean distance metric is a very powerful one for the features studied, other distance metrics have been proposed based on fixed parameter sets, such as the LPC parameters, etc.[3,11] Investigations into the applicability of such distance metrics to the silence-unvoiced-voiced classification problem are currently in progress.

REFERENCES

1. B. S. Atal and L. R. Rabiner, "A Pattern-Recognition Approach to Voiced-Unvoiced-Silence Classification With Applications to Speech Recognition," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-24*, No. 3 (June 1976), pp. 201–212.
2. L. R. Rabiner, M. J. Chang, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-24*, No. 5 (October 1976), pp. 399–418.
3. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-23*, No. 1 (February 1975), pp 67–72.
4. B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., *55* (June 1974), pp 1304–1312.
5. J. Burg, "A New Analysis Technique for Time Series Data," NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands, 1968.
6. J. Makhoul, "New Lattice Methods for Linear Prediction," Proceedings 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing, April 1976, pp 462–465.
7. M. R. Sambur, "Selection of Acoustic Features for Speaker Identification," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-23*, No. 2 (April 1975), pp 176–182.
8. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans Acoust., Speech, and Signal Process., *ASSP-23*, No. 6 (December 1975), pp 552–557.

9. L. R. Rabiner and M. R. Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits," IEEE Trans. Acoust., Speech, and Signal Process., *ASSP-24*, No. 2 (April 1976), pp 170–182.
10. J. L. Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice," Proc. IEEE, *64*, No. 4 (April 1976), pp 405–415.
11. A. H. Gray, Jr., and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoust., Speech, and Signal Process. *ASSP-24*, No. 5 (October 1976), pp 380–391.